BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 1. 75–90
Advance Access publication December 12, 2007
doi:10.1093/bib/bbm059

Biomedical ontologies: a functional perspective

Daniel L. Rubin, Nigam H. Shah and Natalya F. Noy

Submitted: 26th June 2007; Received (in revised form): 7th November 2007

Abstract

The information explosion in biology makes it difficult for researchers to stay abreast of current biomedical knowledge and to make sense of the massive amounts of online information. Ontologies—specifications of the entities, their attributes and relationships among the entities in a domain of discourse—are increasingly enabling biomedical researchers to accomplish these tasks. In fact, bio-ontologies are beginning to proliferate in step with accruing biological data. The myriad of ontologies being created enables researchers not only to solve some of the problems in handling the data explosion but also introduces new challenges. One of the key difficulties in realizing the full potential of ontologies in biomedical research is the isolation of various communities involved: some workers spend their career developing ontologies and ontology-related tools, while few researchers (biologists and physicians) know how ontologies can accelerate their research. The objective of this review is to give an overview of biomedical ontology in practical terms by providing a functional perspective—describing how bio-ontologies can and are being used. As biomedical scientists begin to recognize the many different ways ontologies enable biomedical research, they will drive the emergence of new computer applications that will help them exploit the wealth of research data now at their fingertips.

Keywords: *ontologies; annotation; data analysis*

INTRODUCTION

The e-science era has brought a proliferation in both data and databases, as well as an exponential growth in published literature [1]. Researchers must aggregate and integrate all of this information, and they need tools to enable knowledge discovery in this data-rich paradigm [2]. They have recently begun using ontologies to describe the structure of their complex domains and to relate their data to shared representations of biomedical knowledge. Spurred by the impact and success of the Gene Ontology (GO) [3–5], ontologies have captured the interest of the entire biomedical community. Many groups are creating ontologies, from biologists and bench researchers [6–8] to clinical researchers and clinical practitioners [9–12]. There have been several reviews on biomedical ontology [13–16], highlighting or cataloguing many of the potentially promising ontologies that have appeared, as well as common ontology tools.

There are two complementary perspectives of bio-ontologies: (i) content-oriented view, concerned with the *specific ontologies* being created in biomedicine, and (ii) a functional view, dealing with *how* ontologies can be used to enable a diversity of biomedical applications. The content-oriented view has been addressed well in prior reviews [13–15], describing the activities of individuals and

Corresponding author. Daniel L. Rubin, Stanford Center for Biomedical Informatics Research, Stanford, CA, USA. Tel: 650 723 6979; Fax: 650 725 7944; E-mail: rubin@med.stanford.edu

Daniel L. Rubin is a research scientist at the Stanford Center for Biomedical Informatics Research, and Clinical Assistant Professor of Radiology at Stanford University. He is Scientific Director of the National Center for Biomedical Ontology and a member of the Protégé group. His research focuses on using ontologies to enable intelligent computer applications in biomedicine.

Nigam H. Shah is a research scientist at the Stanford Center for Biomedical Informatics group and member of the National Center for Biomedical Ontology. His research focuses on developing ontology-based approaches to integrate diverse information types for reasoning about biological systems.

Natalya Noy is a Senior Research Scientist at Stanford Center for Biomedical Informatics. She is a member of the Protege group and of the National Center on Biomedical Ontology, where she works on tools for ontology management, including versioning, mapping and modularization of ontologies and on collaborative techniques for ontology development and evaluation.

communities engaged in creating and improving ontologies, projects to accumulate and catalogue ontologies and efforts to critique ontologies and to develop best practices for how ontologies should be created. The functional view focuses on *ontology consumers*—addressing how ontologies can assist *biomedical researchers* find information and interpret their data, as well as providing specific ideas to *ontology creators and curators* for novel ontologies and applications. In this review, we summarize biomedical ontology from this functional perspective, organizing the presentation according to how ontologies are used in biomedical applications.

FUNCTIONAL OVERVIEW OF BIOMEDICAL ONTOLOGIES

A wide variety of artifacts are called 'ontologies' in the biomedical domain, leading to much debate and confusion. The most widely used ontological artifacts are *terminologies*, or *controlled vocabularies* (CVs). A CV provides a list of concepts and text descriptions of their meaning and a list of lexical terms corresponding to each concept. Concepts in a CV are often organized in a hierarchy. Thus, CVs provide a collection of terms that researchers can use for indexing resources, such as records in a database. The GO [3] is the most widely used CV serving biomedical researchers. The GO provides terms for declaring molecular functions, biological processes and cellular components of gene products.

Information models (or data models) are another common ontological artifact. An information model provides an organizing structure to information pertaining to a domain of interest, such as microarray data, and describes how different parts of the information at hand, such as the experimental condition and sample description, relate to each other. In biomedical research, Microarray Gene Expression Object Model (MAGE-OM) is an example of a widely known information model. MAGE-OM, along with the controlled terms that are used to populate the information model is referred to as the Microarray Gene Expression Data (MGED) Ontology [17]. The MGED Ontology is used to describe the minimum information about a microarray experiment that is essential to make sense of the numbers comprising the microarray data.

Finally, there are *ontologies* in the sense of formal representations of knowledge with definitions of concepts, their attributes and relations between them expressed in terms of axioms in some well-defined logic. In biomedical research, several ontologies are striving towards this goal. An example is the Foundational Model of Anatomy (FMA), which is a knowledge resource for anatomy and represents the classes and relationships necessary for the symbolic modeling of the structure of the human body in a form that is understandable to humans and is also navigable, parseable and interpretable by machine-based systems [18].

The range of ontology content and structure is mirrored by the diversity of applications of ontologies from the functional perspective. In this review, we group the uses of ontologies into the following classes of biomedical applications:

- Search and query of heterogeneous biomedical data
- Data exchange among applications
- Information integration
- Natural Language Processing
- Representation of encyclopedic knowledge
- Computer reasoning with data

When describing each group of use, we will select an example ontology to illustrate how that ontology has been used in that area to address the specific biomedical needs.

Search and query of heterogeneous biomedical data

It is challenging to unify diverse data sets in a consistent way when the biological relevance of the same entity—such as association with disease or the involvement in certain processes—is labeled differently in different resources. The language of biomedicine contains many synonymous terms, abbreviations and acronyms that can refer to the same concept. For example, the process of creating glucose is referred to using a variety of synonymous terms, including 'glucose synthesis', 'glucose biosynthesis', 'glucose formation', 'glucose anabolism' and 'gluconeogenesis'. An ontology can provide a single identifier (the class or term identifier) for describing such information for each entity and can store alternative names for that entity through the appropriate metadata. The ontology can thus be used as a controlled terminology to describe biomedical entities in terms of their functions, disease involvement, etc, in a consistent way. In addition, the ontology can be augmented with terminological

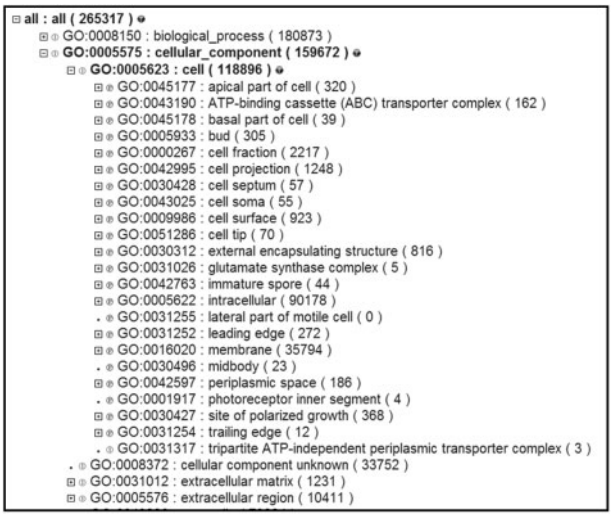

Figure 1: The Gene Ontology. The Gene Ontology as displayed in the Amigo Browser (amigo.geneontology.org/). The cellular component branch of GO is expanded, showing that it comprises a hierarchically organized set of terms describing the components making up cells, with children being related to parent terms via is-a relations. The GO terms are used to provide a controlled terminology for annotating biomedical databases and for creating computable biomedical assertions.

knowledge such as synonymy, abbreviations and acronyms. Ontologies thus enable the community to integrate resources by providing the ability to reliably identify a particular entity or a group of entities based on their biological relevance, such as all proteins associated with cell death.

The gene ontology

The GO [3] is perhaps the canonical example of an ontology created for the primary purpose of providing controlled terms for describing biological entities. Historically, different Model Organism Databases (MODs) described the same functions, biological processes and cell components of gene products using different terms. To enable MODs to describe data unambiguously, the GO Consortium was established to create standard sets of terms for describing biological processes, molecular functions and cellular components of gene products (Figure 1). These terms describe what the gene products do, where they act, and how they perform these activities.

Thus, the GO has enabled all of the MODs to declare functions, processes and cellular-component associations of gene products in an unambiguous manner. To make these assertions, a list of GO terms is associated with each gene product in a process referred to as 'annotation' (Figure 2).

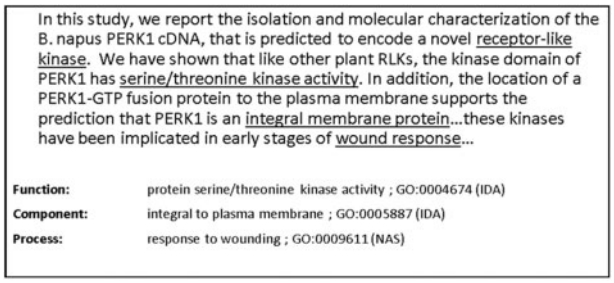

Figure 2: Ontology-based Annotation. An example of use of GO to create annotations based on biomedical literature. In the excerpt from the literature shown, there are several biomedical assertions that are made (underlined text) about the PERK1 gene of *B. napus*. These assertions describe the molecular function (serine/threonine kinase activity), cellular component (integral membrane protein) and biological process (wound response) of PERK1, summarized using the appropriate GO terms from each of the three GO ontologies.

Such ontology-based annotations are highly valuable both for querying databases and for analyzing high throughput data in the following ways:

- **GO-based queries:** researchers can search GO annotations to find all gene products that are involved in particular biological processes (such as all gene products involved in the process of cell death), that have certain molecular functions, or that are located in a specific cellular component.
- **GO-based analysis of high throughput data:** Analysis of GO codes associated with the results of high-throughput data analysis provides biomedical insights into experimental results. The most common task is that of **finding over-represented GO categories** in a list of genes [19]. The most common procedure for this analysis is to count the number of genes with a particular GO annotation (e.g. genes associated with cell death) in the set of significant genes and to compute the probability for finding the number of genes with that particular GO annotation (cell death), assuming the set of significant genes was a random sample from the genome. If this probability is below a certain (usually arbitrary) threshold, then we conclude that the set of genes is 'significant' in either causing or responding to the experimental condition under which the data were collected.

Other ontologies used for describing entities

The Medical Subject Headings (MeSH) is a terminology created by the National Library of

Medicine for indexing the medical literature [20]. MeSH provides a standard set of terms that medical librarians use to describe the main topics covered in papers, such as the species studied, funding source and other attributes. Originally conceived to help librarians carry out literature search, the standard names provided by MeSH have proven useful for augmenting natural language processing methods for text processing, extraction and classification [21–24]. The use of MeSH for providing names for biomedical entities, such as diseases, in these applications is analogous in purpose to the use of GO for providing standard names for biological processes and molecular functions.

The NCI Thesaurus, developed by the National Cancer Institute, integrates molecular and clinical cancer-related information [25]. It provides a controlled terminology that enables researchers to integrate, retrieve, and relate diverse data collected in cancer research. It covers topics such as cancers, findings, drugs, therapies, anatomy, genes, pathways, cellular and subcellular processes, proteins, and experimental organisms. The terminology enables scientists to label experimental results in a standard manner, analogous to GO. NCI Thesaurus enables scientists to link their research findings to disease and molecular patterns [26]. Associating research data with ontology terms also enables efficient search and retrieval of that data, by querying using terms at different levels within the ontology [27] (Figure 3).

RadLex is a controlled terminology for radiology, providing terms for the techniques, findings, and diseases associated with medical images [28]. The Biomedical Informatics Research Network (BIRN) is creating ontologies to provide the necessary names for interrelating concepts contained in images as well as in distributed online databases [29]. The use of ontologies for naming entities in the images permits these projects to unify image data and non-image data, as well as streamline search in large image repositories.

Data exchange among applications

Ontologies can be used to specify how information is organized in biomedical databases as well as how data can be exchanged with other resources. The use of ontology in this context is to represent the information itself, to provide an explicit specification of the terms used to express the biomedical information. The ontology makes explicit the relationships among data types in databases, enabling applications to deduce subsumption among classes. One of the benefits of using ontologies for information models is that the information they describe can be published on the Semantic Web if the ontology is represented in RDF Schema or the Web Ontology Language (OWL) [30]. As a result, the growing arsenals of Semantic Web tools can use, integrate and analyze the data.

MAGE-OM, MAGE-ML and the MGED ontology

Microarrays are a common experimental method being used to measure molecular-level biomarkers for a variety of biological states and medical diseases. The creation of large amounts of microarray data and databases for sharing these data resulted in the need for standards in describing microarray experiments and results. The MIAME standard specifies the minimum information needed to describe a microarray experiment, and the MAGE-OM and markup language MAGE-ML provide a mechanism for standardized representation of microarray data for data exchange [17]. The MAGE-OM declares how the different pieces of information relate to one another, and the MAGE-ML conveys the instance data. The MGED Ontology provides a common terminology and structure for annotating microarray experiments, including the design of the experiment and array layout, the preparation of the biological sample and the methods used to analyze the data and is closely tied to the MAGE-OM.

MAGE-OM is currently the primary information model for describing microarray experiments. MAGE-ML is an XML-based markup language that is derived from MAGE-OM and is used to communicate information about microarray-based experiments among researchers and microarray databases.

BioPAX

Knowledge about biomedical pathways is central to scientific research. There are more than 200 biomedical databases containing information pertinent to pathways. BioPAX is an emerging format for sharing knowledge about pathways. It aims to provide a standard for representing metabolic, biochemical, transcription regulation, protein synthesis and signal transduction pathways [31]. BioPAX does not attempt to provide an ontology of the pathway domain *per se*, but rather a model for pathway data exchange. However, BioPAX is implemented in OWL, so pathways expressed in BioPAX could be seen as ontological representations

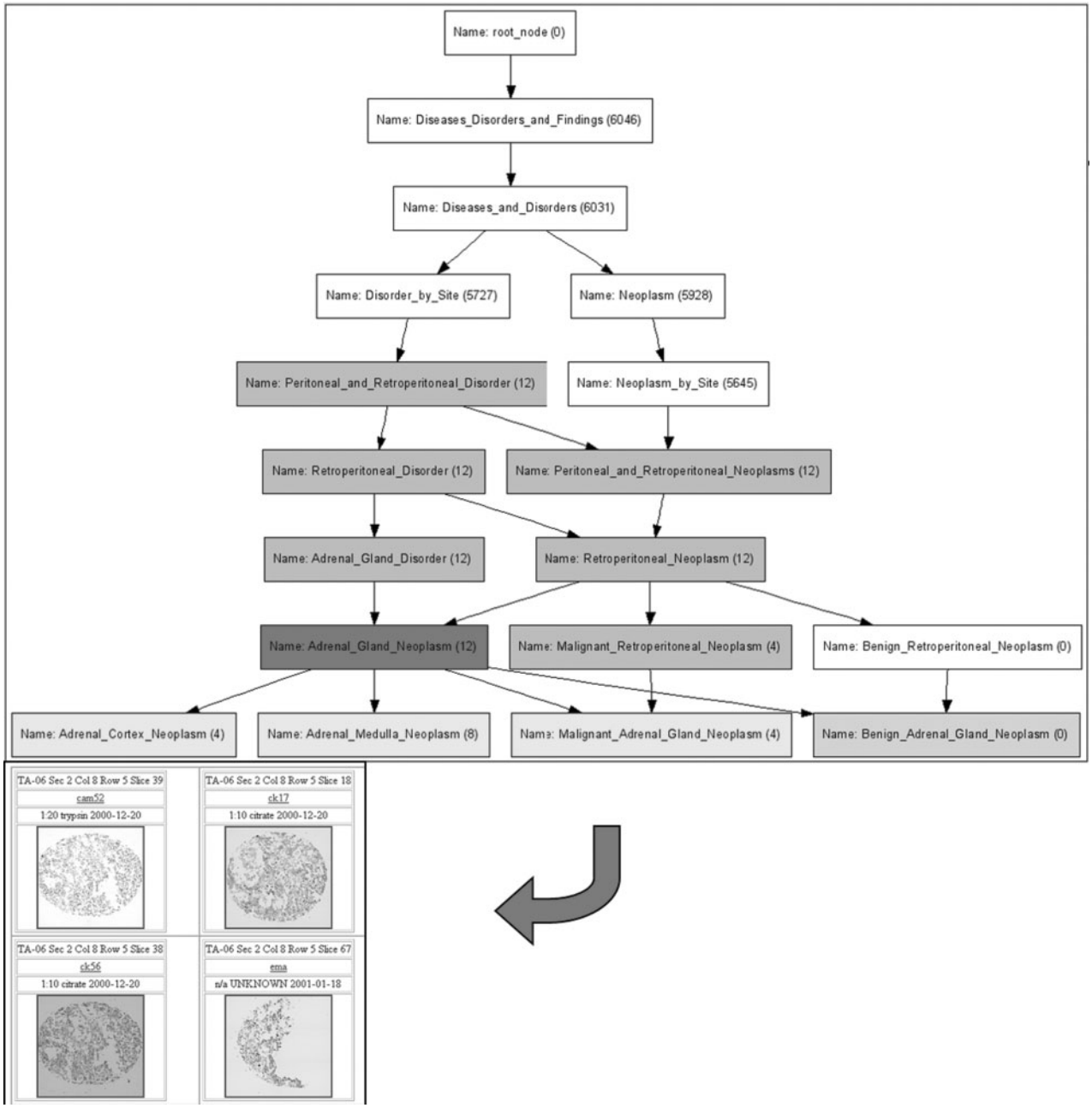

Figure 3: Querying at different levels using an ontology. The figure shows a zoomed in region of the directed acyclic graph view resulting from searching for the term Adrenal gland neoplasm. The red node is the term that has been searched for and then clicked by the user, the yellow nodes are the child terms that have at least one sample in the database assigned to that term, grey nodes are child terms with no corresponding samples in the database and burly-wood nodes are parent terms with less than 50 samples. Samples can be retrieved for the selected node.
Source: Reprinted from Shah, Rubin, Espinosa *et al.*, *BMC Bioinformatics* 2007;8:296, originally published in BioMed Central.

of the corresponding biological pathways and used for inferential analysis.

Pathways in BioPAX are composed of a set of interactions. The top-level BioPAX definition of pathway is general enough to capture the many kinds of pathways used by biologists. Different kinds of pathway information can be represented in BioPAX.

Currently, leading pathway resources such as the Kyoto Encyclopedia of Genes and Genomes

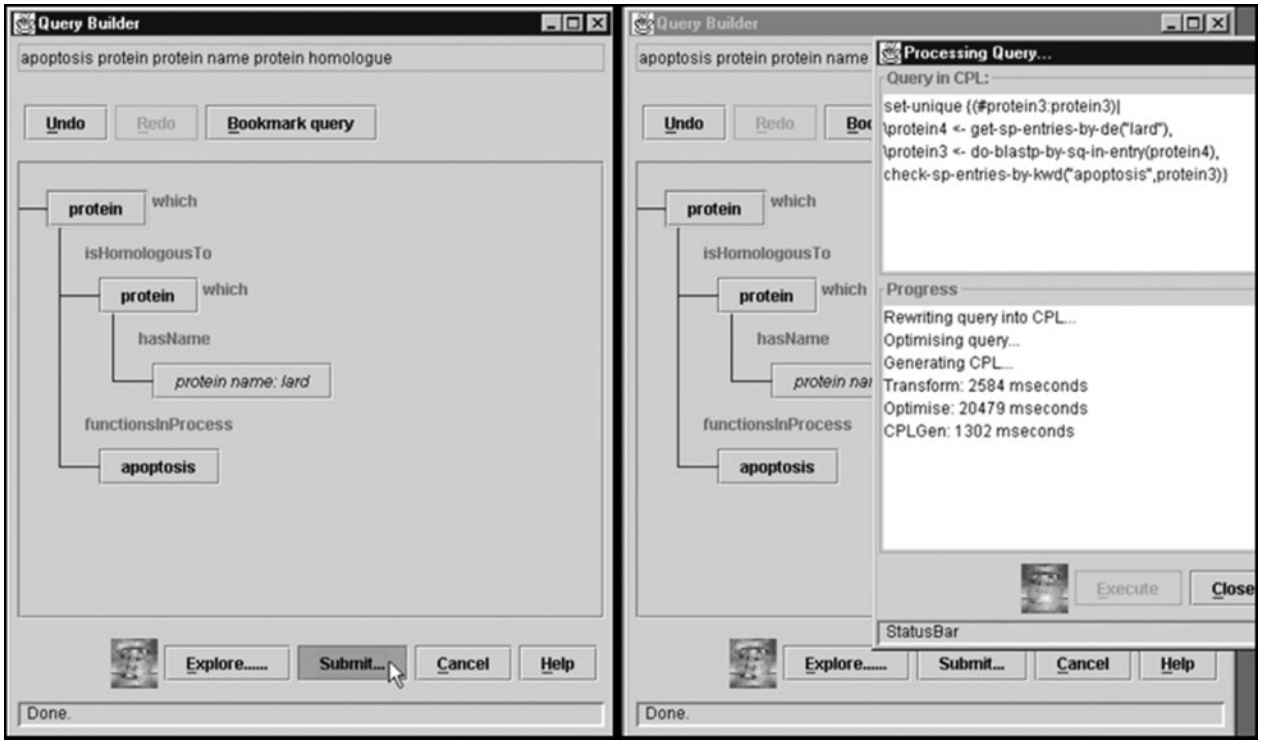

Figure 4: Using ontologies for integrating data resources. Screen shots from the TAMBIS application are shown. A user query is formulated in terms of entities from the TAMBIS ontology (left). In the figure, the query illustrated is: 'find proteins that are homologous to the *lard* protein and functions in the apoptosis biological process'. This query is based on a overarching Concept Model that subsumes all the information models reflected in the databases that the TAMBIS system covers. The Concept Model-based query is translated into appropriate database-specific queries by the application, issued to each appropriate database, and the results collected and returned to the user (right).

(KEGG) [32], BioCyc [33] and Reactome [34] make their data available in BioPAX format and BioPAX viewers are available as additional modules in pathway analysis tools such as PATIKA [35] and Cytoscape [36]. This provides an opportunity to construct unified pathway resources such as the Pathway Knowledge Base project (PKB), enabling querying across different species and across multiple pathway resources simultaneously. It also enables comparison of the degree of complementary across different pathway sources.

Information integration

Ontologies can streamline the process of integrating, accessing and querying data across diverse resources by providing a description of the *contents* of biomedical databases and how the contents of different databases relate to one another. Computer reasoning programs can then use such overarching ontologies to determine the appropriate set of resources to retrieve data required to answer queries that span across resources.

TAMBIS

TAMBIS is a project that aimed to provide transparent access to disparate biological databases and analysis tools, enabling users to access and integrate virtually a wide range of biomedical resources [37]. TAMBIS includes an ontology (the TAMBIS ontology[1]), a knowledge base of biological terminology (the biological Concept Model), a model of the underlying data sources (the Source Model) and a user interface. The Concept Model provides the user with the concepts necessary to construct queries, and shields the user from the details of the various database sources. The Source Model provides a description of the underlying sources and mappings between terms used in the sources and terms in the biological Concept Model. The TAMBIS ontology can be a single access point for multiple biological information sources. Queries are phrased in terms of the ontology, and TAMBIS converts them to query requests to appropriate information sources.

Figure 4 shows a tool that uses the TAMBIS ontology to allow users to formulate a query across a set of diverse biomedical sources, virtually integrating these resources. The tools transforms source-independent, declarative queries formed from terms in the Concept Model into a set of source-dependent, executable procedures.

The query process in the tool shown in Figure 4 proceeds in the following phases:

(i) **Query formulation:** The user formulates a query using terms and relationships in the TAMBIS ontology by constructing a high-level concept describing information of interest using ontology terms. The output of this phase of the application is a source-independent conceptual query.
(ii) **Query transformation:** The tool examines the ontology terms comprising the query to identify biomedical database sources needed to answer query, and it then constructs a query plan tailored to the requirements of each source database.
(iii) **Query execution:** The tool submits the individual queries to the relevant source databases and collects the results, returning them to the user. The TAMBIS developers created wrappers for each source database so the latter can be accessed in syntactically consistent manner.

BIRNLex

Researchers in the Biomedical Informatics Research Network (BIRN) have recently begun using ontologies to integrate image- and non-image data pertaining to the neurosciences [29]. This community is developing BIRNLex,[2] a standardized lexicon of terms for annotating BIRN data sources. The BIRN data sources include structural and functional magnetic resonance imaging (fMRI) databases and multi-scale image databases from mouse models of human neurological disease using MRI, light and electron microscopic imaging. BIRNLex includes terms for neuro-anatomical nomenclature, experimental paradigm names and definitions and some basic imaging and data collection terms. Similar to the data integration applications for TAMBIS, BIRNLex is enabling polling of distributed data sources to find data that fit researchers' query needs (e.g. 'Find all the imaging datasets from schizophrenic subjects with particular clinical test, from every available database registered to BIRN').

BIRN employs a 'mediator architecture' to link multiple databases, each maintaining their specific local schema, into an accessible federated platform. In a model similar to the TAMBIS tool, the mediator in the BIRN databases uses ontologies to relate and integrate the various source databases when processing a user query. The mediator parses the query and subsequently submits database-specific queries to the relevant data sources. BIRN is currently working on enriching the knowledge contained in the ontology [38], migrating the ontology to OWL and creating the necessary classes as well as rich relations that will provide the knowledge necessary for computer reasoning with the integrated data sources [39].

Natural language processing

Natural language processing (NLP) methods—applied to biomedical text in order to extract information—are increasingly using ontologies. The specific role that ontologies play in NLP applications varies according to the expressivity of the ontologies employed for these tasks. At one end of the spectrum, ontologies provide lexicons to recognize named entities or concepts in text. At the other extreme, ontologies guide NLP by providing knowledge models and templates for capturing facts from text.

Textpresso

Textpresso is an ontology-based system for extracting and retrieving specific information from biomedical text [40]. It is a text-mining tool that allows researchers to locate information of interest in the literature. The Textpresso ontology provides a list of standard categories of biological entities commonly sought in literature (such as genes, mutants, pathways, etc) and patterns for relationships that exist among those entities. For example, the relationship 'activate' may exist between ten category types and there are two lexical variants. Textpresso processes the literature, indexing sentences using terms and patterns from this ontology. The power of Textpresso's search engine is manifest in category searches. When searching for a category, the search is restricted only to those terms that populate that category. For example, if one were to do a simple keyword search for 'HSN', all sentences that mention HSN would be retrieved, whereas in a

category search where 'HSN' is restricted to the category of, say, the *C. elegans* gene, only those sentences that mention the HSN gene of *C. elegans* would be retrieved. In a category search using relationships, only those sentences in which the relation (such as activate) is found between the allowable category types are retrieved; just a mention of the text string 'activate' will not be retrieved. Thus, the ontology provides knowledge that expands (or restricts) the search and improves information retrieval.

Geneways

The GeneWays project developed a knowledge model that enables representation of signal-transduction pathways in eukaryotes [41]. Unlike the Texpresso ontology, the GeneWays ontology represents the interactions between molecular substances as well as between substances and processes. The ontology is used to store facts extracted from multiple electronic versions of scientific publications using natural language processing techniques to build a consensus view of signal-transduction pathways.

As the accuracy of information extraction in such systems improve in the future, their performance could rival that of human curation [42]. Ultimately, such systems could enable computer inference applications operating on biomedical literature.

Representation of encyclopedic knowledge

Using ontologies as a source for standardized names is perhaps the simplest use of an ontology, but it does not utilize the expressive power of ontologies for representing knowledge about relationships between the biological entities. Many textbooks describe the components making up living systems (the entities) and how they work and interact with other components (the relations). Describing complex knowledge in texts makes that knowledge accessible to humans, but not to machines. Ontologies are increasingly being used to structure and make explicit encyclopedic biomedical knowledge in a form that is accessible to both researchers and machines.

The foundational model of anatomy

The Digital Anatomist Foundational Model of Anatomy (FMA) [18] is a comprehensive ontology of human anatomy. The FMA contains more than 70 000 entities that describe the elements of canonical human morphology, providing detailed declarative descriptions of anatomic structures. FMA is a *reference ontology* because it specifies canonical knowledge for the domain of anatomy, in the form of a comprehensive set of entities and a large set of relationships (Figure 5). The FMA was created through disciplined representation of the structural organization of the human body in collaboration with anatomists and knowledge engineers with the goal of providing an electronically accessible encyclopedic reference for anatomic knowledge.

The knowledge in FMA can be used in applications that require detailed information about entities beyond their name or their biological association [43, 44]. Software applications that need anatomical knowledge about particular organs can access FMA as a reference and look up entities and their relations in FMA to determine canonical facts about anatomic structures, such as organ composition, continuity and adjacency (Figure 5).

For example, a tool could be created to use anatomic reference knowledge in the FMA to help radiologists interpret imaging studies, informing them about the anatomic structures affected by abnormalities in adjacent organs (Figure 6). Radiological imaging interpretation can be challenging because anatomic knowledge is needed, but access to that information is limited. Anatomy is incompletely visualized in imaging procedures; some anatomic structures are not visible in radiology images due to limited spatial resolution or to individual patient characteristics (Compare Figure 6C and Figure 6B). A tool can use FMA to recognize that small anatomic structures such as the thoracic duct are adjacent to larger, visible structures such as the esophagus (Figure 6A), and inform the radiologist that an abnormality, such as a mass in the esophagus, may be affecting the adjacent thoracic duct, even though the latter is not visible in the radiographic image (Figure 6D). Such detailed anatomic knowledge is useful in informing practitioners about diagnostic possibilities that might be overlooked.

Rubin and colleagues showed that FMA can be useful as a reference knowledge source to predict the anatomic consequences of penetrating injury [45]. The authors developed a software application to deduce all the anatomic structures that could be injured consequent to penetrating trauma—whether those structures were directly in the path of injury or very close to it. In order to perform these inferences, their application used the FMA to find the classes associated with organs that were directly on the path

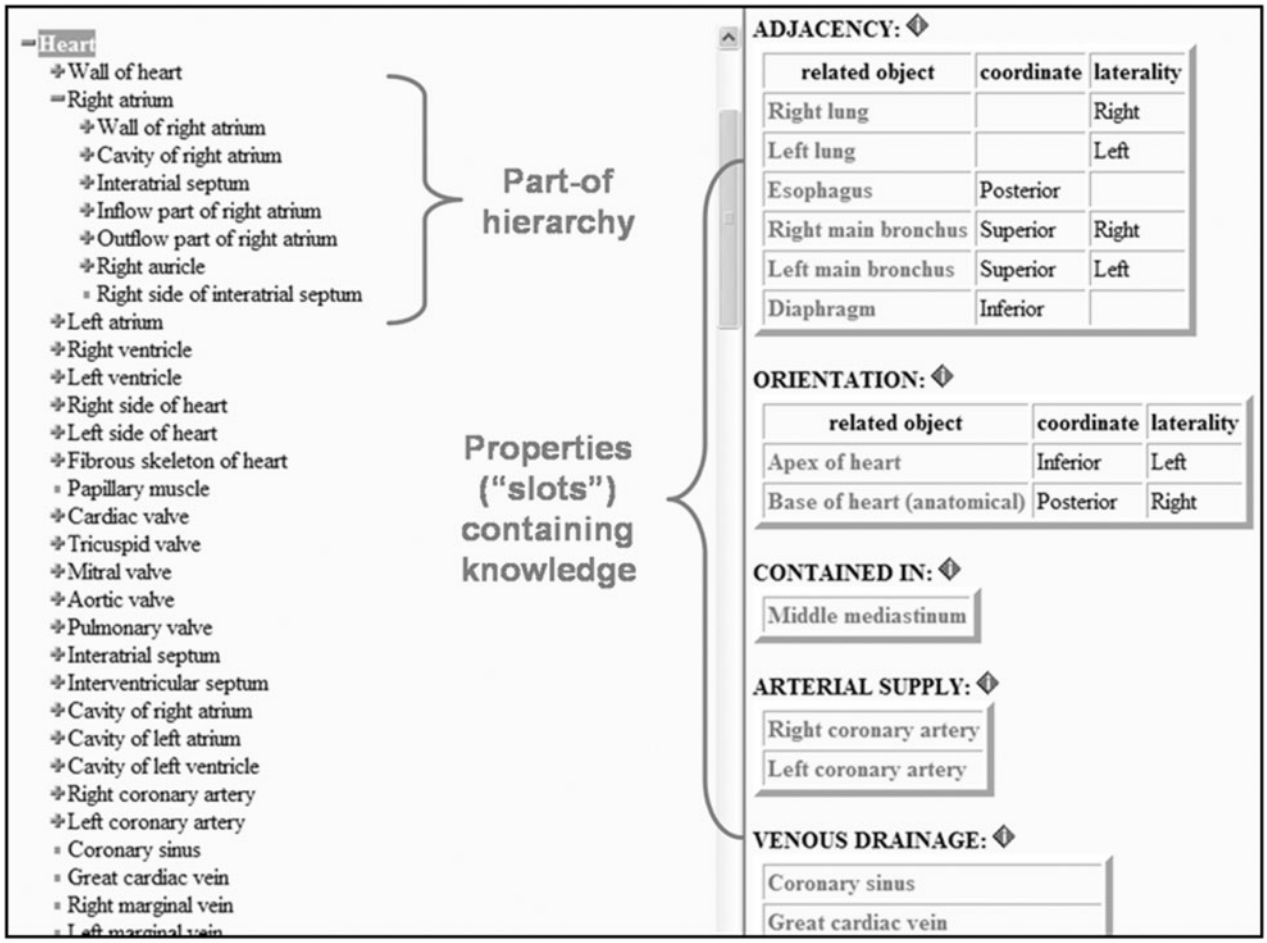

Figure 5: Foundational Model of Anatomy. The FMA is an ontology representing detailed anatomic knowledge. A screen shot of the FMA (accessible by the FM Explorer on the Web at http://fme.biostr.washington.edu:8089/FME/index.html) shows that anatomic knowledge is modeled by specifying a large set of rich relations among the anatomic entities. For example, it can be seen that the heart (left) has many relationships to other entities in the FMA (right), such as adjacency, orientation, containment and vascular supply. Specifically, FMA tells us that the heart is contained in the middle mediastinum, and that it is supplied by left and right coronary arteries.

of injury as well as those adjacent to it, informing caretakers about organs that are injured or potentially injured.

Computer reasoning with data

Computer reasoning is perhaps one of the most compelling advantages ontologies can provide in helping researchers exploit the vast amounts of biomedical knowledge available in electronic form. Computer reasoning encompasses methods that use ontologies to make inferences based on the knowledge they contain as well as any additional contextual information or asserted facts. These methods can help researchers think about what information means in the context of what is already known. Tools can utilize formal methods to query and interpret the information at hand [46].

Hybrow

One of the challenges that computer-reasoning applications can address is integrating current knowledge about biological systems and formulating hypotheses spanning a large number genes and proteins [47]. Currently, it is difficult to determine whether such hypotheses are consistent internally or with data, to refine inconsistent hypotheses and to understand the implications of complicated hypotheses [48].

HyBrow (Hypothesis Browser) is a system for the representation, manipulation and integration of diverse biological data—such as gene expression, protein interactions and annotations—with prior biological knowledge for the purpose of evaluating alternative hypotheses. The goal of Hybrow is to evaluate and rank hypotheses based on user-defined

Figure 6: Using ontologies to enable knowledge-based applications. Among the rich relations contained in the FMA is knowledge about anatomic adjacency—specifications about which anatomic structures are adjacent. Knowledge about adjacency can be used by a computer application to determine which anatomic structures in the vicinity of abnormality may be affected. (**A**) A portion of FMA in Protégé showing that detailed adjacency information is represented; in particular, it can be seen that at the T8 level of the esophagus (highlighted class in left panel), the thoracic duct is located to the right and posterior to the esophagus (value for 'adjacency slot' shown in right panel). (**B**) an image from Visible Human at the T8 level of esophagus showing how adjacency in FMA is established using a relative coordinate system (td = thoracic duct, az = azygous vein, e = eshophagus, a = aorta; this cross section is viewed from below, such that the left side of the patient is on right side of the image). (**C**) An axial Computed Tomography (CT) scan at the same level as (B) showing similar adjacency relations as represented in (A). (**D**) In this CT scan in a different patient from (C), the thoracic duct is not visible, but its presence and location can be deduced from the FMA (A), and this knowledge used to infer that it may be affected by an abnormality (such as a mass) in the adjacent esophagus.

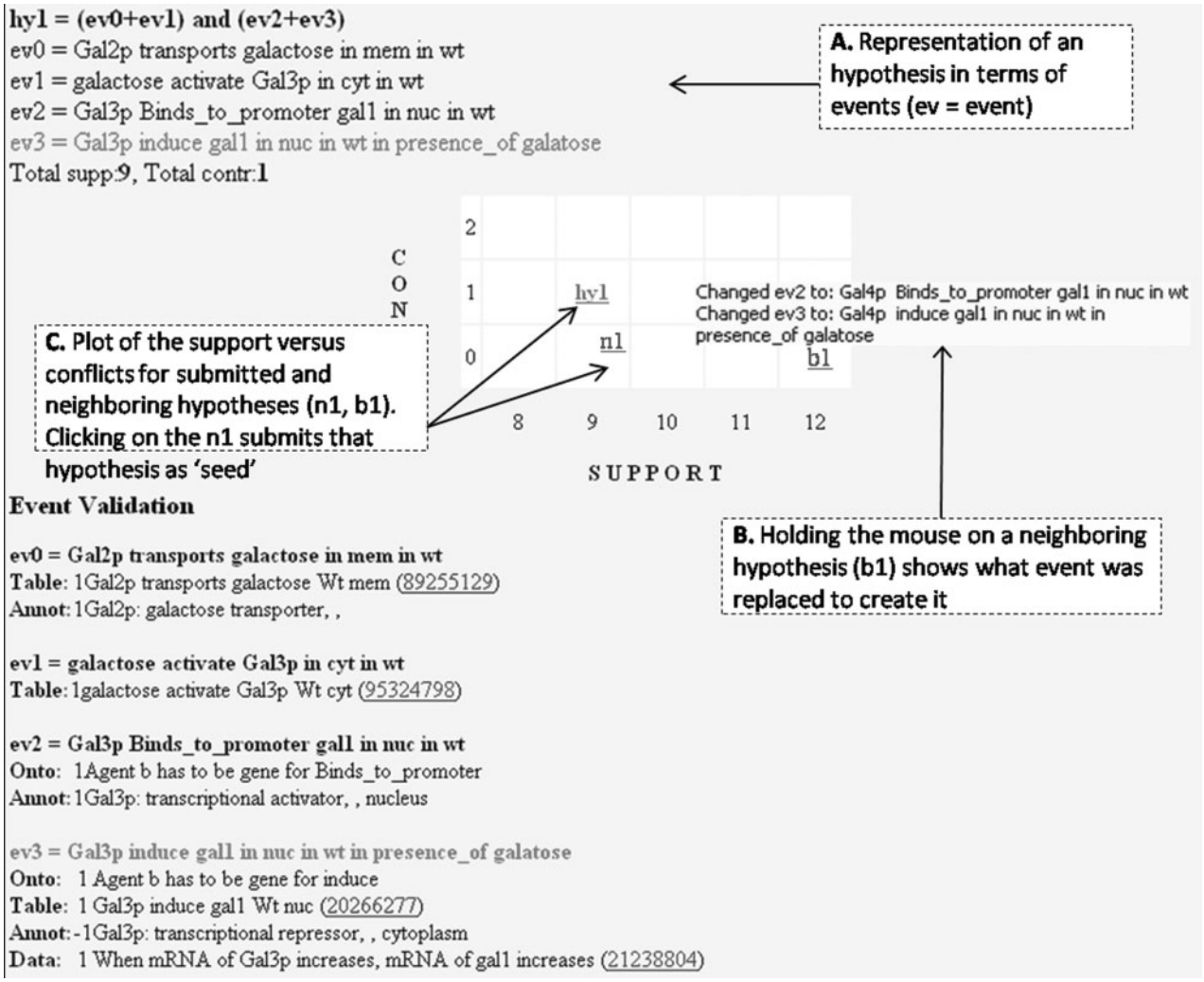

Figure 7: Using ontologies for computer aided reasoning. The figure shows a screen shot from HyBrow's result page that illustrates the evaluation of a simple hypothesis: 'Gal2p transports galactose into the cell at the cell membrane. In the cytoplasm, galactose activates Gal3p. Gal3p binds to the promoter of gall gene and induces its transcription in the presence of galactose'. This hypothesis was decomposed into events (shown in Figure 8A). In its assessment, HyBrow reported support from literature and GO annotation for event number 0 (ev0); support from literature for evl; support from ontology constraints and annotation for ev2 and support from the ontology, literature and data divisions for ev3. HyBrow discovered a conflict for ev3 (marked in light gray) from the annotation rule division since Gal3p is annotated to be primarily in the cytoplasm in presence of galactose. HyBrow then searched for variant events. For ev2 it found an event (Gal4p binds to promoter of gall) with higher support and for ev3 it found the more meaningful event (Gal4p induces gall in nucleus in wt in presence of galactose) with the same support but no conflict. These events were inserted in place of the original events to create a neighboring hypothesis that is better than the original hypothesis. HyBrow is able to present to the user its rankings, explanations for them and references to conflicting and supporting data in a summary page.

'constraints', and to evaluate consistency of hypotheses with all information available [49]. The HyBrow system enables researchers to pose hypotheses. Hybrow then determines whether those hypotheses are consistent with or contradict existing knowledge contained in its knowledge base (Figure 7). HyBrow contains three key components: (i) an event-based ontology for representing hypotheses about biological processes at different levels of detail, (ii) a knowledgebase that stores diverse biological information sources such as gene expression, protein interactions and annotations and (iii) end-user programs to help researchers design hypotheses and evaluate those hypotheses based on the knowledge base and rules of inference.

When a user constructs a hypothesis, HyBrow checks the hypotheses for consistency with the knowledge base, and the support and conflict calls are tallied based upon the logical structure of the hypothesis and presented to the user in a Web interface (Figure 7) [27].

Currently, HyBrow represents knowledge about the Galactose metabolism in yeast (GAL system) [50]. HyBrow's knowledge base stores the existing knowledge about the GAL system. The knowledge base accommodates available literature, curated primarily from YPD [51] at a coarse level of resolution. The knowledge base was populated by manual curation using loading forms like the EcoCyc database [25] as well as PERL scripting to access the existing public repositories (such as the Saccharomyces Genome Database) to retrieve desired information.

Other ontologies and computer reasoning applications

The FMA has been used in a reasoning application to deduce the physiological consequences of injury to the arteries supplying the heart [52]. The authors represented the knowledge about which arteries supply different regions of the heart in an OWL ontology. Connectivity among different arterial segments was specified via continuity relations, and the concepts of arterial occlusion and heart ischemia were defined. A computer reasoning service posed the problem of inferring heart injury as a classification problem, based on asserting arterial injuries, classifying the resulting ontology and reading off newly-classified anatomical entities that reflected the inferred heart injuries [52].

Ontologies also provide knowledge for inference in decision support applications. Decision support applications inform practitioners on the preferred practice or optimal decision given the specific contexts. The ATHENA decision support system implements medical guidelines for high blood pressure in patients [53]. The system helps physicians to manage patients with high blood pressure and recommends guideline-concordant choices of drug therapy. The ATHENA ontology specifies eligibility criteria, risk stratification, blood pressure targets, relevant associated diseases and preferred drugs within each drug class. The system allows clinical experts to customize the ontology to incorporate new evidence or to reflect local interpretations of guideline ambiguities.

FUTURE DIRECTIONS

In this review, we have surveyed biomedical ontologies from a functional perspective—in terms of how they are used in applications. This approach is helpful to those coming into the field to get a sense for the spectrum of potential use cases and to recognize opportunities for ontology to help with their particular use case. This perspective focuses on the *consumers* of ontologies—individuals who care about the ways ontologies can be used to enable research, discovery and health care.

The proliferation in ontologies has also created opportunities for new research for *producers* of ontologies and tools, as well as for the ontology field in general. The community of ontology users will increasingly need tools to help them to find, reconcile and relate the growing number of biomedical ontologies. A number of tools and services for this purpose are already being developed. For example, Swoogle (http://swoogle.umbc.edu/) is a search engine that automatically collects ontologies available on the Web. Users can search for ontologies by entering keywords that Swoogle matches to labels of classes and properties. Swoogle ranks the results based on the information provided by links between ontologies. The Ontology Lookup Service from the European Bioinformatics Institute provides a centralized query interface for ontologies in the Open Biomedical Ontology (OBO) format [54]. BioPortal, created by the National Center for Biomedical Ontology, provides a virtual ontology library [55] where users can submit their ontologies in a variety of ontology formats. BioPortal organizes ontologies according to a set of categories (such as anatomy, genomics, development, etc), enabling users to find groups of ontologies of interest (Figure 8) as well as to visualize them (Figure 9). BioPortal users will be able to rate ontologies, comment on how appropriate ontologies are for specific tasks and how well they cover their target domain.

Furthermore, as the field of biomedical ontologies expands, ontologies inevitably will cover overlapping domains, and researchers will need to relate them to one another. Researchers are developing a plethora of tools for identifying relations between concepts in different ontologies automatically or semi-automatically [56]. These tools perform with varying effectiveness [57] and none of them are perfect. In the future, we envision that tools, like BioPortal, will integrate the results of automatic algorithms with mappings

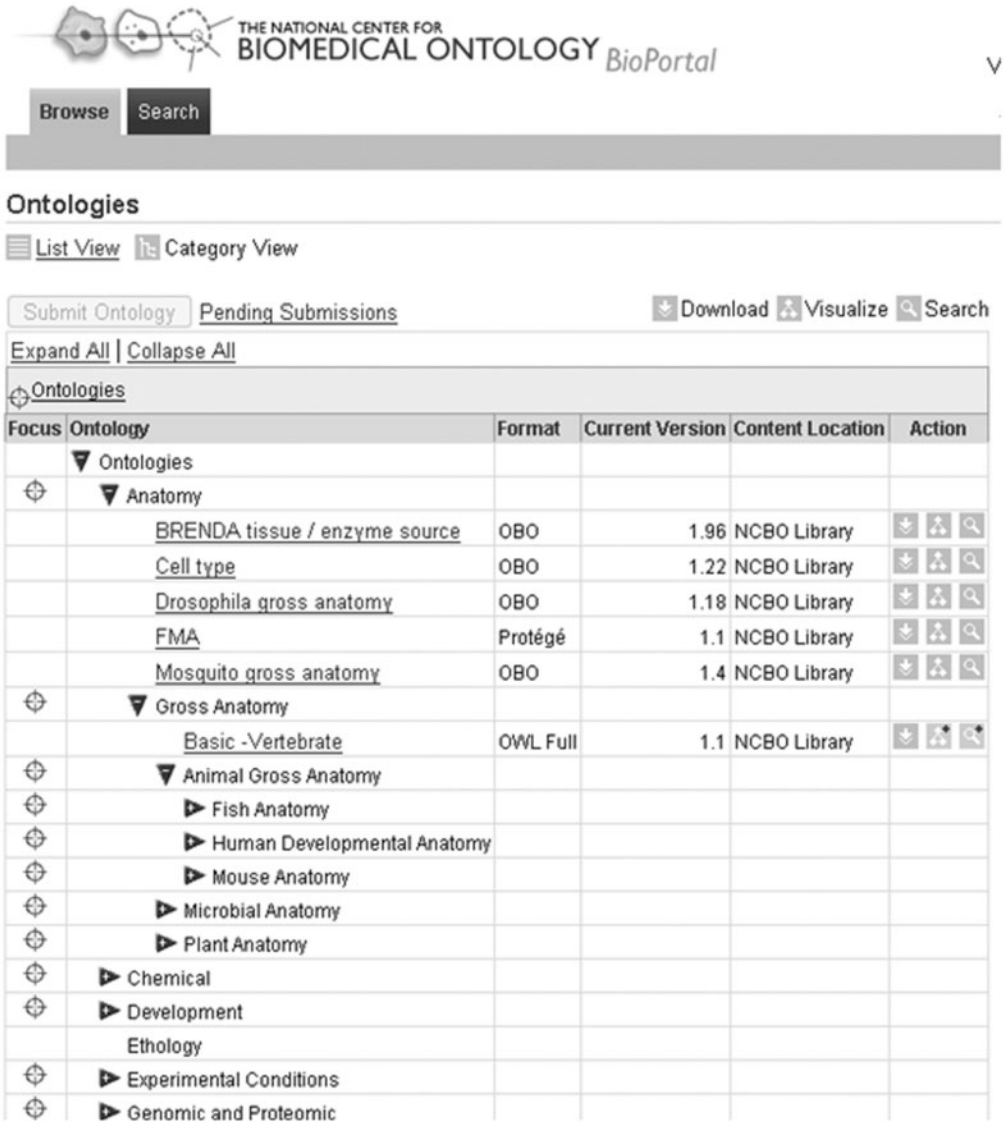

Figure 8: BioPortal Ontology Library displayed in the category view, showing a taxonomy of ontologies, organized according to the type of ontology.

developed by a community of domain experts in an open way.

Another future direction for ontology research is in developing metrics for ontology quality and in creating tools to enable the user community to evaluate ontology quality. As ontologies get used more widely in different applications, their users may want to report feedback, to point to errors, and to request additions and extensions from ontology authors [58, 59]. Tools such as BioPortal will provide the community with the ability to submit feedback on the quality and utility of particular ontologies. For example, people browsing ontologies will be able to suggest that a class should be moved to a different part of the ontology, that it should be split into two or more classes, or that it should be renamed and so on. Such structured feedback could include information about who created the note and when, so that the ontology developers can establish trust relationships in this user feedback, enabling them to determine those comments to which they should pay particular attention.

CONCLUSIONS

The number of biomedical researchers interested in biomedical ontology has been rapidly expanding, as

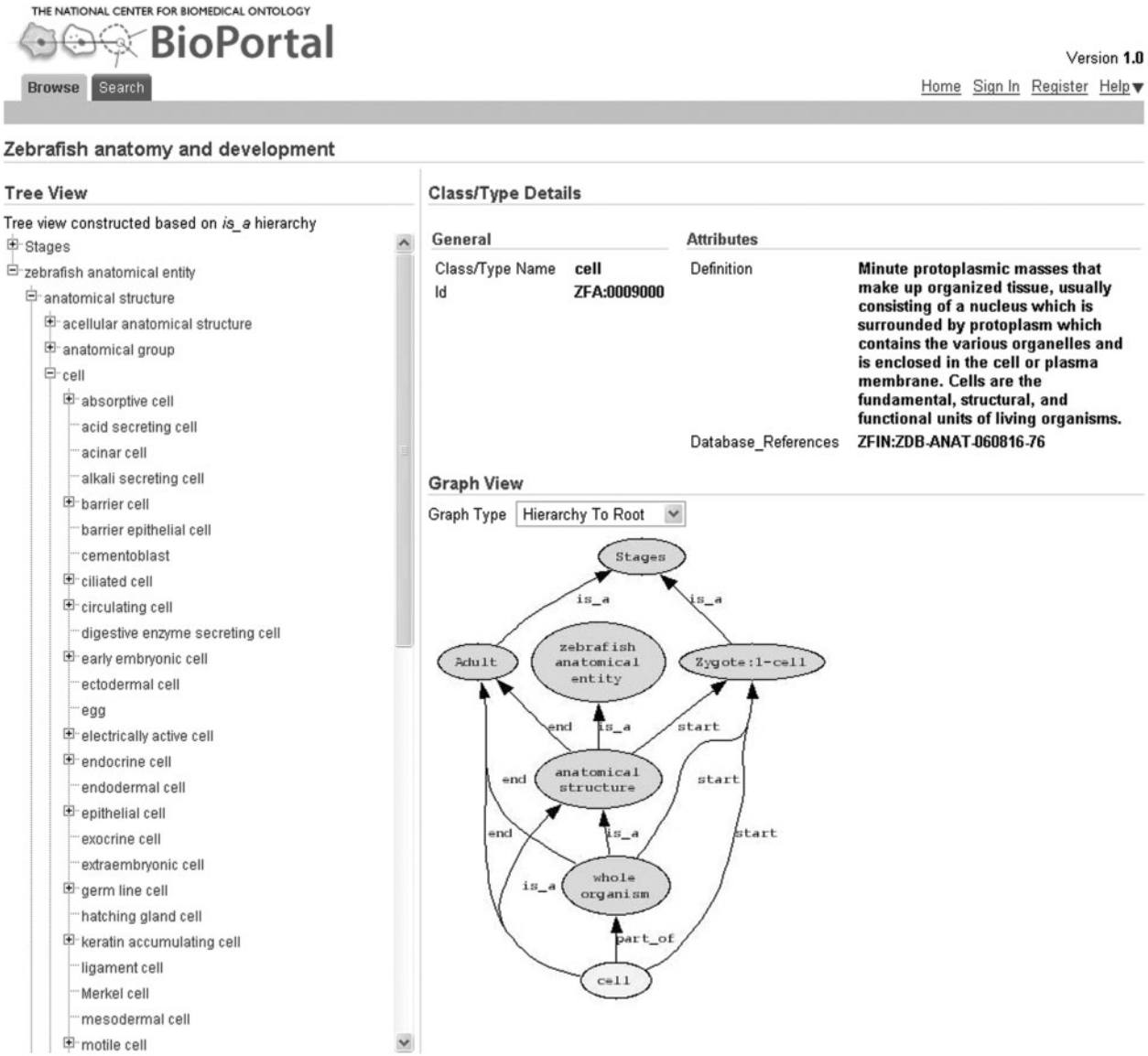

Figure 9: BioPortal: Ontology access and visualization. In BioPortal, ontologies are shown both as an expandable tree (left) as well as a local neighborhood graph (right; selected class is highlighted in light gray).

have the number of ontologies. This growth in interest and content has both enabled and fragmented the field. For those not already highly knowledgeable about biomedical ontology, it can be helpful to organize ones thinking about them from a functional perspective—how can ontologies be used to enable biomedical research. For those interested in participating in developing ontologies, the paradigms for creating them is evolving, bringing new challenges related to coordinated development and maximally benefiting from reuse. At the same time, new tools are appearing to meet the challenges and to enable biomedical researchers to benefit fully from the growing ontology resources.

Key Points

- There is a growing community interested in using and producing biomedical ontologies.
- The diversity of ontologies and their content can bewilder those not already deeply familiar with the field; it is helpful to consider bio-ontologies from a functional perspective.
- Ontologies are used in biomedicine for describing biological entities, specifying information models, enabling Natural Language Processing, specifying data exchange formats and semantics of data for information integration as well as for providing reference encyclopedic knowledge and enabling computer reasoning with biomedical data.
- There will likely be continued growth in biomedical ontologies as well as new tools and paradigms for people to work with them.

Notes

1 http://www.ontologos.org/%5COntology%5CTAMBIS.htm
2 http://132.239.132.249:8080/xwiki/bin/view/+BIRN-OTF-Public/About+BIRNLex

Acknowledgements
This work was supported by the National Center for Biomedical Ontology, under roadmap-initiative grant U54 HG004028 from the National Institutes of Health.

References

1. Hey T, Trefethen AE. Cyberinfrastructure for e-Science. *Science* 2005;**308**:817–21.
2. Fedoroff N, Racunas SA, Shrager J. Making biological computing smarter. *The Scientist* 2005;20–21.
3. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
4. Harris MA, Clark J, Ireland A, *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;**32**:D258–61.
5. Lewis SE. Gene Ontology: looking backwards and forwards. *Genome Biol* 2005;**6**:103.
6. Drysdale RA, Crosby MA, Gelbart W, *et al*. FlyBase: genes and gene models. *Nucleic Acids Res* 2005;**33**:D390–5.
7. Sprague J, Clements D, Conlin T, *et al*. The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res* 2003;**31**:241–3.
8. Rhee SY, Dickerson J, Xu D. Bioinformatics and its applications in plant biology. *Annu Rev Plant Biol* 2006;**57**:335–60.
9. Dybkaer R. An ontology on property for physical, chemical, and biological systems. *APMIS Suppl* 2004;**117**:1–210.
10. Huq SZ, Karras BT. A proposed ontology for online healthcare surveys. *AMIA Annu Symp Proc* 2003;304–9.
11. Ma J. Building an Ontology-driven database for clinical immune research. *AMIA Annu Symp Proc* 2006;1018.
12. Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence–based practice: content specification and evaluation using the competency decomposition method. *J Biomed Inform* 2004;**37**:108–19.
13. Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. *Methods Inf Med* 2006;**45**(Suppl 1):124–35.
14. Yu AC. Methods in biomedical ontology. *J Biomed Inform* 2006;**39**:252–66.
15. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform* 2006;**7**:256–74.
16. Rojas I, Ratsch E, Saric J, *et al*. Notes on the use of ontologies in the biochemical domain. *In Silico Biol* 2004;**4**:89–96.
17. Ball CA, Brazma A. MGED standards: work in progress. *Omics* 2006;**10**:138–44.
18. Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform* 2003;**36**:478–500.
19. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005;**21**:3587–95.
20. Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Jama* 1994;**271**:1103–8.
21. Bodenreider O. Using UMLS semantics for classification purposes. *Proc AMIA Symp* 2000;86–90.
22. Hersh W, Hickam DH, Haynes RB, *et al*. Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature. *Proc Annu Symp Comput Appl Med Care* 1991;808–12.
23. Rubin DL, Thorn CF, Klein TE, *et al*. A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. *J Am Med Inform Assoc* 2005;**12**:121–9.
24. Suomela BP, Andrade MA. Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics* 2005;**6**:75.
25. Hartel FW, de Coronado S, Dionne R, *et al*. Modeling a description logic vocabulary for cancer research. *J Biomed Inform* 2005;**38**:114–29.
26. Sioutos N, de Coronado S, Haber MW, *et al*. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007;**40**:30–43.
27. Shah N, Rubin D, Supekar K, *et al*. Ontology-based annotation and query of Tissue Microarray Data. *AMIA Annu Symp Proc* 2006;709–13.
28. Langlotz CP. RadLex: a new method for indexing online educational materials. *Radiographics* 2006;**26**:1595–7.
29. Martone ME, Gupta A, Ellisman MH. E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat Neurosci* 2004;**7**:467–72.
30. Horrocks I. An ontology language for the semantic Web. *IEEE Intell Syst* 2002;**17**:74–5.
31. BioPax-Consortium. BioPAX: Biological Pathways Exchange. http://www.biopax.org/(Dec 2006, date last accessed).
32. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**:D354–7.
33. Karp PD, Ouzounis CA, Moore-Kochlacs C, *et al*. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005;**33**:6083–9.
34. Joshi-Tope G, Gillespie M, Vastrik I, *et al*. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005;**33**:D428–32.
35. Demir E, Babur O, Dogrusoz U, *et al*. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* 2002;**18**:996–1003.
36. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
37. Stevens R, Baker P, Bechhofer S, *et al*. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* 2000;**16**:184–5.

38. Gupta A, Ludascher B, Grethe JS, *et al*. Towards a formalization of disease-specific ontologies for neuroinformatics. *Neural Netw* 2003;**16**:1277–92.
39. Li C, Maryann M, Amarnath G, *et al*. OntoQuest: exploring ontological data made easy. *Proceedings of the 32nd international conference on Very large data bases - Volume* ;**32**. Seoul, Korea: VLDB Endowment, 2006.
40. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2004;**2**:e309.
41. Rzhetsky A, Koike T, Kalachikov S, *et al*. A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics* 2000;**16**:1120–8.
42. Rodriguez-Esteban R, Iossifov I, Rzhetsky A. Imitating manual curation of text-mined facts in biomedicine. *PLoS Comput Biol* 2006;**2**:e118.
43. Brinkley JF. Structural informatics and its applications in medicine and biology. *Acad Med* 1991;**66**:589–91.
44. Rosse C, Mejino JL, Modayur BR, *et al*. Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base. *J Am Med Inform Assoc* 1998;**5**:17–40.
45. Rubin DL, Dameron O, Bashir Y, *et al*. Using ontologies linked with geometric models to reason about penetrating injuries. *Artif Intell Med* 2006;**37**:167–76.
46. Gifford DK. Blazing pathways through genetic mountains. *Science* 2001;**293**:2049–51.
47. Kuchinsky A, Graham K, Moh D, *et al*. Biological storytelling: a software tool for biological information organization based upon narrative structure. *Advanced Visual Interfaces*. New York: ACM, 2002.
48. Karp PD. Pathway databases: a case study in computational symbolic theories. *Science* 2001;**293**:2040–4.
49. Racunas SA, Shah NH, Albert I, *et al*. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics* 2004;**20**:i257–64.
50. Racunas SA, Shah N, Fedoroff NV. A Contradiction-based framework for testing gene regulation hypotheses. *IEEE Bioinformatics*. Palo Alto, California: IEEE Computer Society, Stanford University, 2003.
51. Proteome. Yeast Proteome Database. http://www.proteome.com/YPDhome.html (14 April 2002, date last accessed).
52. Rubin DL, Dameron O, Musen MA. Use of description logic classification to reason about consequences of penetrating injuries. *AMIA Annu Symp Proc* 2005;649–53.
53. Goldstein MK, Hoffman BB, Coleman RW, *et al*. Implementing clinical practice guidelines while taking account of changing evidence: ATHENA DSS, an easily modifiable decision-support system for managing hypertension in primary care. *Proc AMIA Symp* 2000;300–4.
54. Cote RG, Jones P, Apweiler R, *et al*. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* 2006;**7**:97.
55. Rubin DL, Lewis SE, Mungall CJ, *et al*. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *Omics* 2006;**10**:185–98.
56. Noy NF. Semantic integration: a survey of ontology-based approaches. *Sigmod Record* 2004;**33**:65–70.
57. Euzenat J, Mochol M, Shvaiko P, *et al*. Results of the ontology alignment evaluation initiative 2006. In: Shvaiko P, Euzenat J, Noy N, *et al*, (eds). *Proc. 1st ISWC 2006 International Workshop on Ontology Matching*. Trento, IT: Universita degli Studi di Trento.
58. Noy NF, Guha RV, Musen MA. User ratings of ontologies: who will rate the raters? *Proceedings of the AAAI 2005 Spring Symposium on Knowledge Collection from Volunteer Contributors*. Trento, IT: Universita degli Studi di Trento, 2005 Available at: http://www.stanford.edu/~natalya/papers/SS505NoyN.pdf.
59. Noy NF, Rubin DL, Musen MA. Making biomedical ontologies and ontology repositories work. *IEEE Intell Syst* 2004;**19**:78–81.